SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA – PIBIC

ESTIMAÇÃO DE "PITCH" E AVALIAÇÃO ESPECTROGRÁFICA VOCAL: UMA PARCERIA ENTRE A ENGENHARIA ELÉTRICA E A FONOAUDIOLOGIA NA AVALIAÇÃO DO SINAL VOCAL.

Área do conhecimento: Engenharia Elétrica
Linha de Pesquisa: Automação Inteligente

Relatório Final
Período da bolsa: 01 de Agosto de 2017 a 31 de Julho de 2018

Este projeto foi desenvolvido com bolsa de iniciação científica

PIBIC/CNPq

Orientador: Jugurta Rosa Montalvão Filho
Co-orientadora: Ariane Damasceno Pellicani
Autor: Jônatas Cruz Santos

Sumário

1 Introdução

A produção vocálica humana é um processo complexo, que envolve a cooperação entre diversos mecanismos do corpo humano. De uma maneira simplificada, podemos entender a voz como o resultado da modificação de uma produção sonora laríngea, por meio de ressonâncias, i.e., "a voz é o som produzido pela vibração das pregas vocais, modificado pelas cavidades situadas abaixo e acima dela" [1]. A região iniciada na glote e encerrada nos lábios, pela qual a vibração das pregas vocais é modificada, é denominada trato vocal.

O termo *pulso glotal* se refere ao sinal produzido pela vibração das pregas vocais. O processo dessa vibração ocorre com a sua abertura e fechamento de maneira cíclica (ou quase cíclica), de modo que a taxa de repetição desse movimento define a frequência fundamental (parâmetro objetivo de um sinal periódico do qual o *pitch* é uma medida perceptual análoga) da voz. A voz resultante emitida é moldada pelo ajuste do trato vocal, cuja disposição momentânea define as frequências em que ocorrerão ressonâncias. Essas frequências são denominadas formantes [2].

Na fonoaudiologia, o processo de avaliação vocal é um importante procedimento cujo objetivo é conhecer e estudar o comportamento e a saúde vocal de um dado indivíduo, com o intuito de compreender o comportamento vocal [1] e avaliar a sua qualidade. Diversos procedimentos podem ser realizados para a avaliação e o diagnóstico das disfonias, como por exemplo a avaliação otorrinolaringológica, a análise acústica e a análise perceptivo-auditiva [3].

A avaliação otorrinolaringológica geralmente faz uso da endoscopia para a obtenção de um diagnóstico médico para o distúrbio vocal [1]. Apesar de se tratar de um método eficiente, esse processo de avaliação é invasivo ao paciente e necessita de um procedimento médico que se deseja evitar. Assim, a avaliação otorrinolaringológica é empregada com mais frequência após a realização de outras análises por parte de um fonoaudiólogo, caso seja necessário.

A avaliação da voz por meio da análise perceptivo-auditiva toma como base a percepção do fonoaudiólogo e é o método mais comumente utilizado para a avaliação das disfonias e da qualidade vocal. Isso ocorre devido ao fato de que essa forma de avaliação procura levar em conta fatores biológicos, psicológicos e/ou sociais [4]. Entretanto, essa forma de avaliação apresenta muita subjetividade e dependência da experiência e do treinamento do fonoaudiólogo avaliador.

A análise acústica da voz utiliza de técnicas do processamento digital de sinais para a estimação dos parâmetros acústicos da voz, o que garante objetividade à análise. Atualmente, a análise acústica da voz é considerada complementar à análise perceptivo-auditiva, pelo fato de não possibilitar a percepção da limitação social de uma voz [1]. Entretanto, diversos estudos indicam que a avaliação por meio de parâmetros acústicos da voz é capaz de oferecer maior objetividade e precisão na identificação de problemas vocais [5], tornando possível a detecção de pequenas alterações que são impossíveis de serem notadas pelo ouvido humano [6].

Este projeto se apoiou numa cooperação entre a fonoaudiologia e a engenharia eletrônica da UFS, no estudo e modelagem matemática/computacional do trato vocal e do pulso glotal, tendo como perspectiva principal o auxílio à avaliação vocal e à interação entre fonoaudiólogos e pacientes. Para isso, foram propostas representações visuais de amostras de voz, com base em parâmetros extraídos destas, que representam de maneira significativa ao fonoaudiólogo as alterações vocais, sendo

também de fácil entendimento ao seu paciente. Dessa forma, essa representação pode se tornar uma alternativa válida para uma melhor comunicação clínica e também uma nova ferramenta que auxilie o fonoaudiólogo em diagnósticos e tratamentos.

No decorrer do projeto foram explorados diferentes métodos para a estimação de *pitch* [7], de frequências formantes [8], de pulso glotal [9], entre outros parâmetros acústicos da voz, que foram empregados para a caracterização da qualidade vocal e para a elaboração da representação intuitiva da voz proposta.

Neste relatório são apresentados todos os procedimentos teóricos-experimentais desenvolvidos durante este projeto de pesquisa. O texto encontra-se disposto de maneira que na seção 2 são apresentados os objetivos do trabalho, seguidos pela descrição da metodologia na seção 3. Na seção 4 os resultados obtidos são apresentados e discutidos, sendo as considerações finais feitas na seção 5. Em seguida, as perspectivas futuras do trabalho são apresentadas na seção 6 e, após as referências bibliográficas, são descritas outras atividades, complementares, desenvolvidas.

2 Objetivos

Este trabalho possui o propósito de investigar o uso de parâmetros acústicos da voz como ferramentas para a avaliação vocal e para a comunicação entre fonoaudiólogo e paciente. De uma maneira específica, os seguintes objetivos foram propostos:

1. Estudar e implementar computacionalmente os métodos para a estimação da frequência fundamental (f_0). Em particular, foram estudados os estimadores clássicos voltados a sinais vocálicos humanos [7][10].

2. Estudar e implementar métodos para a estimação de pulso glotal e trato vocal a partir de sinais acústicos [9][11].

3. Propor representação gráfica intuitiva para a voz com base em parâmetros do pulso glotal e/ou do trato vocal.

4. Sugerir avaliação automática da qualidade vocal em escalas perceptuais-auditivas para os parâmetros vocais (como rouquidão, aspereza, soprosidade, entre outros) utilizando medidas objetivas da voz.

3 Metodologia

No decorrer deste projeto, foram estudados os aspectos teóricos e foram implementados os métodos computacionais clássicos para estimação dos parâmetros propostos. Tendo em vista os objetivos apresentados, o trabalho realizado pode ser dividido em quatro âmbitos: estimação de frequência fundamental; estimação de pulso glotal e trato vocal; avaliação automática da qualidade vocal em escalas perceptuais-auditivas; representação gráfica da voz.

Para a implementação dos métodos computacionais utilizados, o processo de produção vocal humano foi modelado, de maneira simplificada, como um processo de filtragem linear, no qual o sinal do pulso glotal é aplicado na entrada de um filtro que representa o trato vocal, gerando o sinal da voz como saída (Figura 1).

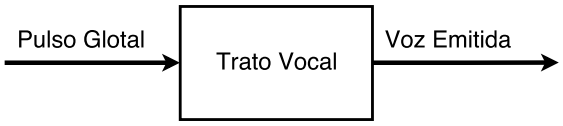

Figura 1. Modelo digital simplificado para sinais de voz utilizado nesse trabalho

3.1 Estimação de frequência fundamental

Para a estimação da frequência fundamental de um sinal, são utilizados métodos capazes de encontrar a repetição cíclica, ou quasi-cíclica, nele existente. Atualmente, diversos algoritmos na literatura se propõem a realizar a estimação do f_0 [7]. Foi utilizado para executar essa tarefa o método baseado na autocorrelação do sinal, um dos mais simples métodos de estimação de f_0. Dado um sinal digitalizado $x[k]$, a função de autocorrelação empírica $\phi[k]$ de um sinal pode ser definida como segue [2]:

$$\phi[k] = \sum_{m=-\infty}^{\infty} x[m]x[m+k] \tag{1}$$

em que se pode inferir que o valor máximo da função de autocorrelação é atingido quando $k = 0$. Também se faz notável que, para sinais periódicos, a função de autocorrelação do sinal apresenta a propriedade de manter-se periódica com o mesmo período do sinal. Isso indica que para $k = nP$ (em que P é o período e n é um número inteiro qualquer), $\phi[0] = \phi[k]$, sugerindo, assim, que a cada ciclo o valor máximo da função de autocorrelação é atingido. Assim, o método para a estimação de f_0 utiliza essas propriedades para encontrar o período P do sinal e, consequentemente, a frequência fundamental do sinal.

3.2 Estimação de Pulso Glotal e Trato Vocal

Pulso glotal e trato vocal de um sinal de voz carregam em si informações relevantes para o entendimento da formação da fala, para a síntese de voz e para o estudo da qualidade vocal. No pulso glotal temos a excitação produzida pelas cordas vocais de maneira pura e no trato vocal encontram-se representadas as ressonâncias que modificam o sinal glótico, gerando a voz como resultado.

Para a estimação dos sinais de pulso glotal e trato vocal foram implementados dois procedimentos de acordo com diferentes metodologias. Primeiramente, foi implementado um procedimento de acordo com a metodologia da análise-pela-síntese [2], que fora inicialmente proposta. Posteriormente, foi implementado o *Iterative Adaptative Inverse Filtering* (IAIF) [12], um método clássico para a estimação do pulso glotal que utiliza predição linear [13] e filtragem inversa [14], e estima o trato vocal no processo. Ambos os procedimentos são descritos em detalhes nos tópicos seguintes.

3.2.1 Análise-pela-síntese

O método da análise-pela-síntese consiste em sintetizar um sinal de voz utilizando modelos paramétricos para o pulso e para o trato e realizar ajustes aleatórios iterativos. Esses ajustes devem ocorrer de maneira que o erro entre o sinal sintetizado e o sinal original da voz seja minimizado, de forma que o sinal sintetizado equivalha, aproximadamente, ao original.

O processo implementado para a estimação do pulso glotal e do trato vocal tem seu início com estimação da f_0, de acordo com a metodologia proposta anteriormente na subseção 3.1. O valor estimado é, então, utilizado para a síntese de um trem de pulsos glotais com a mesma frequência do sinal.

Foi projetado um filtro IIR truncado por um filtro FIR [15], inicialmente com ressonâncias nas frequências de 500 Hz, 1500 Hz e 2500 Hz (3 pares de polos complexos conjugados). Na entrada desse filtro foi aplicado um trem de impulsos, com frequência equivalente à f_0 do sinal, utilizando o impulso digital para simular o pulso glotal, resultando em um sinal sintetizado, com a mesma duração do sinal original, pelo processo de filtragem. A análise-pela-síntese é, portanto, realizada de acordo com a seguinte sequência de passos:

1. É selecionado um trecho do sinal com duração menor que um período, sendo o trecho denominado $X[n]$
2. De igual modo, o trecho equivalente ao mesmo intervalo no sinal sintetizado é selecionado e denominado $Y[n]$
3. Calcula-se o erro quadrático médio entre $Y[n]$ e $X[n]$
4. Os 3 pares de polos são perturbados com pequenas variações aleatórias em seus módulos e suas fases (mantendo a simetria dos pares complexos conjugados)
5. Com os novos valores dos polos o sinal $Y[n]$ é sintetizado novamente e outra vez é calculado o erro quadrático médio entre os dois sinais. Caso o erro seja menor que o erro anterior, os novos polos são mantidos. Caso contrário, descartam-se os polos
6. Retorna-se ao passo 4, repetindo o procedimento por uma determinada quantidade de iterações ou até ser atingido o critério de parada

Posteriormente, o mesmo procedimento foi realizado utilizando, entretanto, o modelo de Rosenberg [16] para a síntese do trem de pulsos glotais. Esse modelo paramétrico representa um único pulso glotal de acordo com a seguinte expressão [2]:

$$g[n] = \begin{cases} \frac{1}{2}\left(1 - cos\left(\frac{\pi n}{N_1}\right)\right), & 0 \leq n \leq N_1 \\ cos\left(\frac{\pi(n-N_1)}{2N_2}\right), & N_1 \leq n \leq N_1 + N_2 \\ 0, & \text{para os demais casos} \end{cases} \tag{2}$$

em que N_1 representa o número de amostras na fase de abertura e N_2 o número de amostras na fase de fechamento, como ilustrado na Figura 2. Nessa etapa, o pulso glotal de Rosenberg também sofre modificações aleatórias para o ajuste.

3.2.2 Predição linear

A predição linear, ou LPC (*Linear Predictive Coding*), consiste na estimação do valor de uma amostra $x[n]$ do sinal a partir das k amostras anteriores (LPC de ordem k), como representado a seguir [17]:

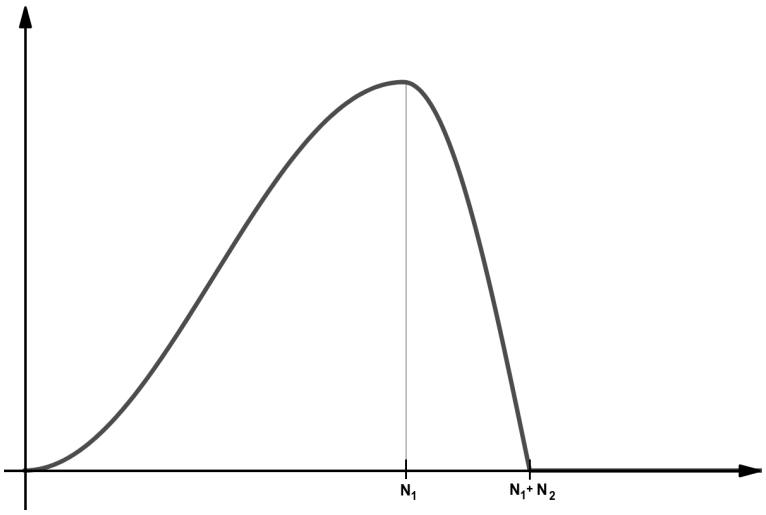

Figura 2. Pulso Glotal de Rosenberg

$$x[n] = \sum_{m=1}^{k} a_m x[n-m] + r[n] \tag{3}$$

em que r[n] representa a perturbação aleatória independente do sinal x[n].

Assim, cada elemento de x poderá ser representado como uma combinação linear entre os k elementos anteriores. É possível representar (3) pelo seguinte produto de matrizes:

$$\begin{bmatrix} x[k+1] \\ x[k+2] \\ \vdots \\ x[N] \end{bmatrix} = \begin{bmatrix} x[1] & x[2] & \dots & x[k] \\ x[2] & x[3] & \dots & x[k+1] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-k] & x[N-k+1] & \dots & x[N-1] \end{bmatrix} \begin{bmatrix} a_k \\ a_{k-1} \\ \vdots \\ a_1 \end{bmatrix} \tag{4}$$

$$Y = MC \tag{5}$$

em que (5) equivale a (4), i.e.:

$$Y = \begin{bmatrix} x[k+1] \\ x[k+2] \\ \vdots \\ x[N] \end{bmatrix}, \quad M = \begin{bmatrix} x[1] & x[2] & \dots & x[k] \\ x[2] & x[3] & \dots & x[k+1] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-k] & x[N-k+1] & \dots & x[N-1] \end{bmatrix} \; e \; C = \begin{bmatrix} a_k \\ a_{k-1} \\ \vdots \\ a_1 \end{bmatrix}$$

Assim, C é obtida de acordo com o método dos mínimos quadrados, utilizando o cálculo da pseudo-inversa [18] de M:

$$C = M^{\dagger} Y \tag{6}$$

Com os coeficientes a_m encontrados (para m de 1 a k), assumindo que (3) representa um filtro ressonador (e assumindo uma entrada aleatória e independente no sistema) uma função de transferência pode ser calculada para o sinal analisado. Para isso, a transformada Z [15] é aplicada em (3), resultando na seguinte expressão:

$$H(z) = \frac{1}{1 - \sum_{m=1}^{k} a_m z^{-m}} \tag{7}$$

A estimação do contorno espectral é obtida a partir do cálculo do módulo de $H(z)$ no intervalo de frequências de 0 até a metade da frequência de amostragem. O comportamento espectral de $H(z)$, de acordo com a ordem do LPC, ocorre como o exemplificado pela Figura 3, em que se pode notar também como a ordem do preditor influencia no detalhamento da resposta espectral.

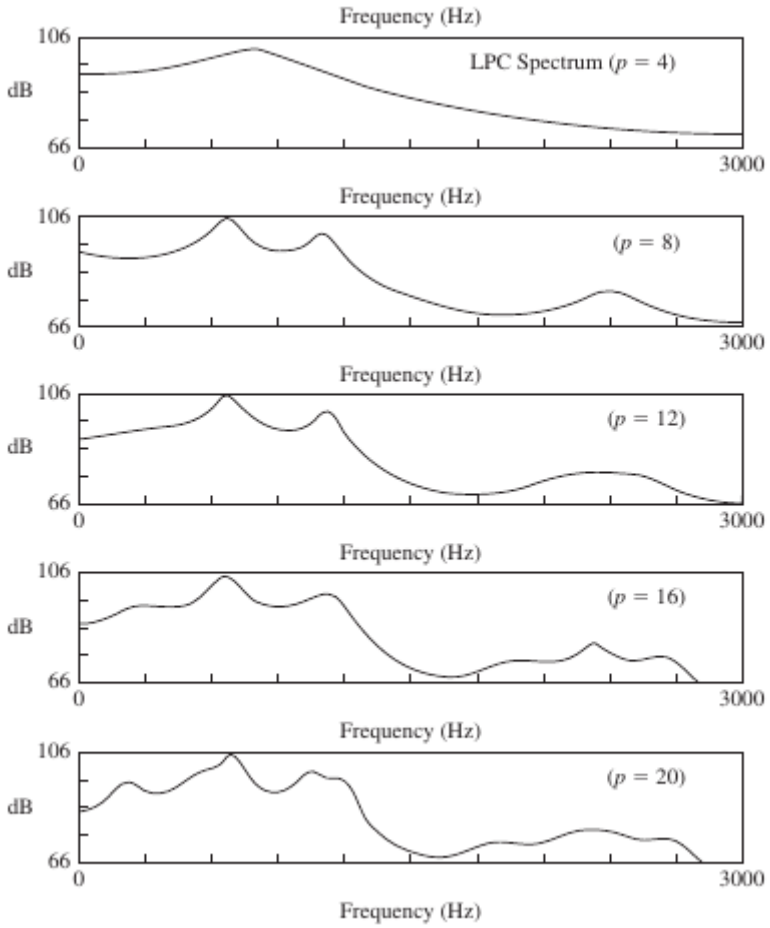

Figura 3. Variação espectral da estimação por LPC do sinal, em função da ordem p. Aqui os eixos verticais correspondem ao módulo de $H(z)$
(Fonte: Rabiner e Schafer [19])

3.2.3 Filtragem Inversa

O processo de filtragem inversa consiste em filtrar o sinal $x[n]$, removendo as influências dos polos da função de transferência em (7) ($H(z)$), utilizando um filtro $G(z)$ que pode ser representado como segue:

$$G(z) = H(z)^{-1} \tag{8}$$

3.2.4 Iterative Adaptative Inverse Filtering

O IAIF [12] é um método que emprega predição linear e filtragem inversa utilizando uma estrutura iterativa para estimar o pulso glotal. Para sua implementação, o trato vocal (visto como um filtro digital linear) no modelo digital da voz, representado na

Figura 1, é decomposto, considerando a ação da radiação ocorrida nos lábios, que possui o efeito similar ao de um diferenciador aproximado.

O procedimento proposto pelo IAIF consiste em realizar uma filtragem passa-altas em um sinal de voz, $s[n]$, e aplicar, no sinal resultante, o procedimento ilustrado na Figura 4. Primeiramente é realizada a estimação da contribuição glotal, $H_g(z)$, obtida por predição linear de ordem 1, seguida pela remoção dessa por meio de filtragem inversa. Em seguida, é estimada a contribuição do trato vocal, $H_v(z)$, utilizando predição linear de ordem v, sendo essa removida por filtragem inversa, seguida da eliminação do efeito da radiação dos lábios por meio de um integrador estimando assim o pulso glotal. Por fim, esse procedimento é executado mais uma vez, utilizando, entretanto, ordem g na estimação da contribuição glotal.

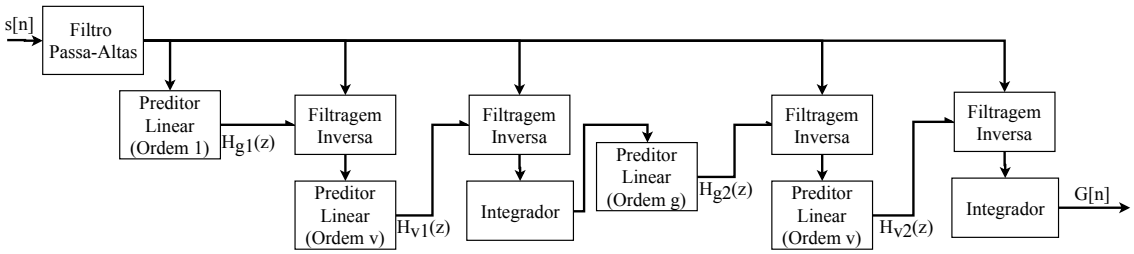

Figura 4. Diagrama de blocos do IAIF

3.3 Avaliação automática da qualidade vocal

Existem diversas escalas propostas para a avaliação vocal empregadas na análise perceptivo-auditiva da voz. Para a avaliação clínica e o estudo das disfonias, utiliza-se com maior frequência as chamadas escalas ordinais e escalas visuais analógicas [20]. Nas escalas ordinais, são atribuídos valores específicos para determinadas características numa escala de n pontos, isto é, dentro de um intervalo discreto de 0 a $n-1$ (em que $n \in N$ e determina a ordem da escala). Nas escalas visuais analógicas, entretanto, são utilizadas linhas, para determinadas características, com um dado comprimento (normalmente $100mm$), nas quais o avaliador indica o ponto na linha correspondente ao grau da característica avaliada [1]. Assim, o valor resultante é dado pelo comprimento do início da linha até o ponto marcado. A escala GRBASI é um exemplo de escala ordinal e a escala do CAPE-V um exemplo de escala visual analógica. Neste trabalho foram utilizadas vozes avaliadas segundo a escala GRBASI.

3.3.1 Escala GRBASI

A escala GRBASI é uma escala ordinal utilizada para a análise perceptivo-auditiva. Originalmente nomeada GRBAS, é uma escala japonesa proposta em 1981 por Hirano [21] como um método simples de avaliação das disfonias. A escala proposta por Hirano é utilizada para a avaliação de 5 itens específicos que dão o nome à escala, a saber **Grau Geral da Disfonia** (*Grade* - G), **Rugosidade** (*Roughness* - R), **Soprosidade** (*Breathiness* - B), **Astenia** (*Asthenia* - A) e **Tensão** (*Strain* - S). Cada uma dessas características da GRBAS é avaliada em uma escala ordinal de 4 pontos, de acordo com a seguinte classificação:

- 0 = Normal

- 1 = Leve ou Discreto
- 2 = Moderada
- 3 = Severa

Somente em 1996 foi proposta a inserção da avaliação da instabilidade vocal à escala GRBAS por Dejonckere et al [22] de modo que a partir de então a escala passou a ser denominada GIRBAS ou GRBASI. As definições das características avaliadas pela escala GRBASI são apresentadas a seguir [1].

- **G** - Representa a impressão geral que se tem do grau de disfonia em uma voz. É utilizada para identificar a disfonia por completo.
- **R** - Engloba em si os conceitos de rouquidão, aspereza, crepitação e bitonalidade. Seu rótulo destaca sua irregularidade e a característica "não lisa" da voz.
- **B** - Representa a presença de uma espécie de chiado, ocasionado por uma abertura muito larga entre as cordas vocais durante a fonação [23], possibilitando um escape de ar na glote.
- **A** - Trata-se de uma espécie de fraqueza vocal caracterizada por uma energia vocal reduzida e uma baixa definição nos componentes harmônicos da voz.
- **S** - Caracterizado por uma frequência fundamental mais aguda que o normal e ruídos de alta frequência. Passa uma "impressão de estado hiperfuncional"[1].
- **I** - Representa a flutuação da qualidade vocal durante o tempo da fonação [22] e a flutuação na frequência fundamental.

3.3.2 Procedimento para a avaliação automática

Buscando conciliar as avaliações perceptivo-auditiva e acústica, foi proposto um método de avaliação automática da qualidade vocal, de acordo com a escala GRBASI, que utilizasse parâmetros acústicos para classificar uma voz de acordo com rótulos atribuídos por fonoaudiologos.

Foi utilizada uma base de dados[1] contendo 79 amostras de voz avaliadas de acordo com a escala GRBASI nas características GRB, disponibilizadas por Alvear et. al [24]. Os sinais da base foram janelados no tempo, utilizando janela retangular [15]. Dessa forma cada trecho dos sinais foi processado, sendo estimados os seguintes parâmetros:

- **Frequência Fundamental**: A frequência fundamental foi estimada, seguindo a metodologia apresentada na subseção 3.1, para os trechos dos sinais. Dessa forma, cada sinal teve como parâmetros a média e o desvio padrão da f_0.
- **Formantes**: Foram utilizadas as formantes dos sinais para cada um de seus trechos. A estimação das formantes foi implementada pela detecção dos picos no contorno espectral (obtido por predição linear como proposto na subseção 3.2). Para cada sinal, foi utilizada a média e o desvio padrão das quatro primeiras formantes.

[1]Disponível em: www.atic.uma.es/index_atic.html

Tabela 1. Codificação One-Hot

Rótulo Original(G, R ou B)	Rótulo Codificado
0	[1 -1 -1 -1]
1	[-1 1 -1 -1]
2	[-1 -1 1 -1]
3	[-1 -1 -1 1]

Esses 10 parâmetros foram utilizados para a avaliação automática da qualidade vocal por meio de um classificador linear [25], i.e., a avaliação automática é realizada por meio de uma combinação linear dos parâmetros. Foi utilizada uma codificação *one-hot* modificada para os rótulos. Nessa codificação, é gerado um vetor preenchido por -1 e um único 1 numa posição referente ao rótulo atribuído, de acordo com o representado na Tabela 1. Assim, o classificador é implementado de acordo com a seguinte expressão:

$$t_{n,\mathtt{r}} = \sum_{k=1}^{10} a_{k,\mathtt{r}} c_{n,k} + erro_{n,\mathtt{r}},\ \forall\ 0 \leq \mathtt{r} \leq 3 \tag{9}$$

em que, para a n-ésima amostra da base, $t_{n,r}$ indica o valor da posição referente ao rótulo r individual da avaliação (G, R, ou B), de acordo com a codificação *one-hot* (ver Tabela 1), a_k é um coeficiente da combinação linear e $c_{n,k}$ é o k-ésimo parâmetro utilizado.

Devido ao pequeno tamanho da base, foi proposto o uso de validação cruzada *leave-one-out*. Esse método consiste em remover uma amostra $n = l$ da base para a estimação dos coeficientes $a_{k,r}$, seguido pelo teste dos coeficientes estimados na amostra removida, alternando o valor de l. Por fim, o total de acertos na classificação é contabilizado.

Dada uma amostra removida l, para $1 \leq l \leq 79$, podemos representar (9) em forma matricial, definindo:

$$T_{l,r} = \begin{bmatrix} t_{1,\mathtt{r}} \\ \vdots \\ t_{l-1,\mathtt{r}} \\ t_{l+1,\mathtt{r}} \\ \vdots \\ t_{79,\mathtt{r}} \end{bmatrix},\ A_{l,r} = \begin{bmatrix} a_{1,\mathtt{r}} \\ \vdots \\ a_{l-1,\mathtt{r}} \\ a_{l+1,\mathtt{r}} \\ \vdots \\ a_{10,\mathtt{r}} \end{bmatrix} e\ C_l = \begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,10} \\ \vdots & \vdots & \ddots & \vdots \\ c_{l-1,1} & c_{l-1,2} & \dots & c_{l-1,10} \\ c_{l+1,1} & c_{l+1,2} & \dots & c_{l+1,10} \\ \vdots & \vdots & \ddots & \vdots \\ c_{79,1} & c_{79,2} & \dots & c_{79,10} \end{bmatrix}$$

de forma que os coeficientes a_k podem ser estimados com o menor erro possível, utilizando pseudo-inversão [18], de acordo com a metodologia dos mínimos quadrados, como representado a seguir:

$$A_{l,\mathtt{r}} = {C_l}^{\dagger} T_{l,\mathtt{r}} \tag{10}$$

Assim, com $A_{l,\mathtt{r}}$ estimado, a classificação para a amostra l é dada por:

$$\hat{t_{l,\mathtt{r}}} = \mathcal{D}\left(\begin{bmatrix} c_{l,1} & c_{l,2} & \dots & c_{l,10} \end{bmatrix} \times A_{l,\mathtt{r}}\right) \tag{11}$$

em que $\hat{t_l}$ é o rótulo estimado para a amostra l e $\mathcal{D}(\ )$ é uma função responsável por receber o resultado do produto e quantizá-lo, de acordo com a codificação *one-hot*.

Isso é feito localizando a posição j do maior elemento do vetor resultante e, assim, atribuir 1 à essa posição na saída ($t_{l=j} = 1$) e -1 às demais posições ($t_{l\neq j} = -1$). A comparação entre $\hat{t_l}$ e t_l determina o acerto ou erro da estimação.

3.4 Representação gráfica da voz

Foi proposta uma representação ilustrativa da voz, levando em conta a sensibilidade natural humana ao reconhecimento de expressões faciais. Para isso, foi implementada uma interface baseada em desenhos de faces parametrizados por características acústicas da gravação de uma vogal sustentada. Dessa forma, possíveis irregularidades na voz interferem nas expressões faciais desenhadas.

A partir do sinais de voz da base disponibilizadas por Alvear et. al [24], ao longo do tempo foram estimados os seguintes seis parâmetros acústicos:

p_1: Desvio padrão da f_0 durante a sustentação da vogal

p_2: Desvio padrão da potência da primeira harmônica do sinal (componente na frequência f_0)

p_3: Desvio padrão da potência das 19 harmônicas seguintes do sinal (componentes nas frequências $2f_0$, $3f_0$, ... $20f_0$)

p_4: Desvio padrão da potência total do sinal

p_5: Proporção entre a potência média da primeira harmônica e a potência média das 19 harmônicas seguintes

p_6: Proporção entre a potência média do sinal e a potência média das 20 primeiras harmônicas

em que, no desenho esquemático de um rosto (como ilustrado na Figura 5): p_1 e p_2 ajustam, respectivamente, a curvatura e a elevação da sobrancelha esquerda; p_3 e p_4 ajustam, respectivamente, a curvatura e a elevação da sobrancelha direita; p_5 e p_6 ajustam, juntos, a curvatura da boca.

É esperado que esses parâmetros representem bem a qualidade de uma voz. Foram utilizadas as médias dos parâmetros de dezenas de amostras de voz avaliadas como G=0 para definir o rosto ideal, ilustrado na Figura 5.

Assim, quando esses parâmetros são utilizados para processar novas amostras de voz, uma nova representação facial é gerada com aspectos visuais determinados pelos parâmetros extraídos. O aspecto visual das faces correspondentes aos sinais de voz foi avaliado subjetivamente.

4 Resultados e discussões

Podemos observar na Figura 6 o resultado da estimação da frequência fundamental, por autocorrelação, em um sinal de voz utilizado como exemplo, no qual há uma emissão da vogal /a/ sustentada.

A análise-pela-síntese foi aplicada para o mesmo sinal de exemplo cuja f_0 foi estimada por autocorrelação (Figura 6). O sinal obtido utilizando o impulso como pulso glotal encontra-se representado na Figura 7. Na Figura 8 é ilustrado o sinal

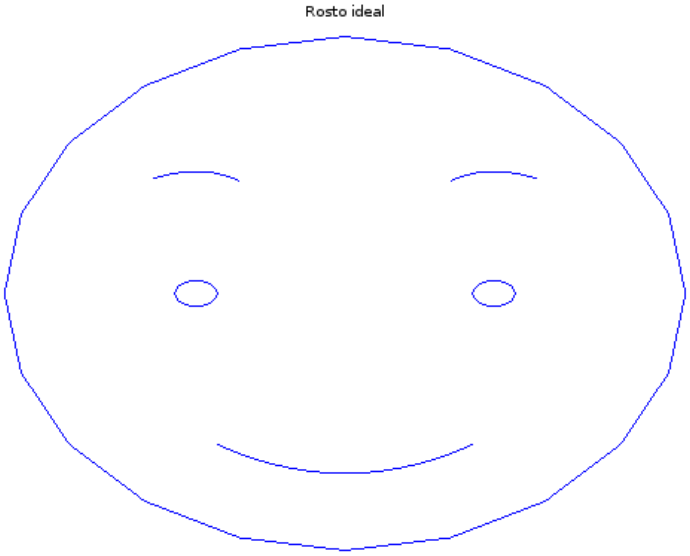

Figura 5. Rosto Ideal

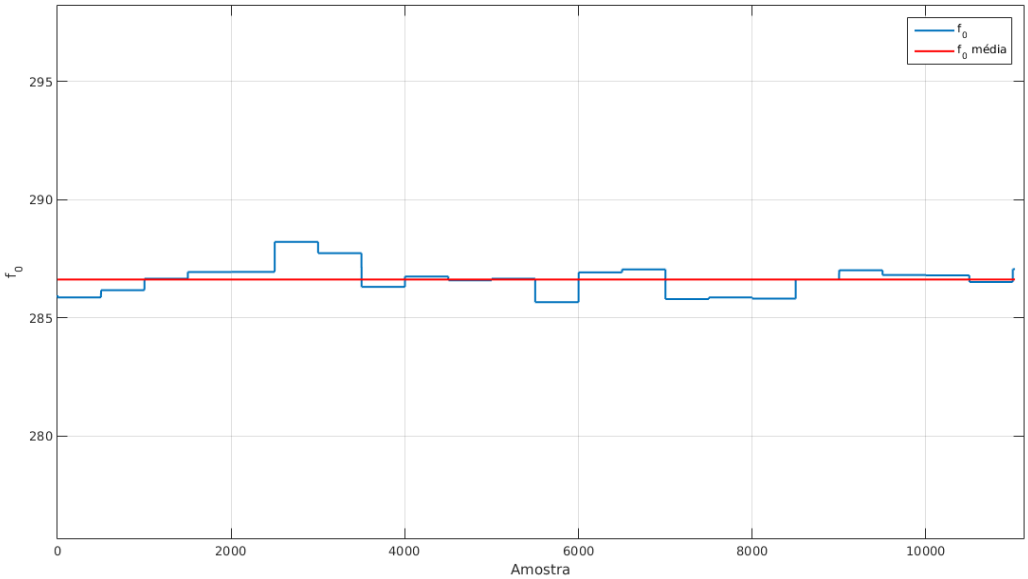

Figura 6. Exemplo aplicado em intervalos de um segundo de voz, com taxa de amostragem de 11025 amostras/s.

obtido ao utilizar o pulso glotal de Rosenberg para a análise. O pulso ajustado a partir do pulso glotal de Rosenberg, para esse exemplo, se encontra apresentado na Figura 9.

Os resultados obtidos utilizando a análise-pela-síntese não se mostraram satisfatórios para a estimação do pulso glotal e do trato vocal. A discrepância entre os formatos de onda do sinal sintetizado e do sinal original evidenciam a inadequação do método implementado.

Na Figura 10 encontram-se representados o sinal de exemplo utilizado anteriormente para a estimação da frequência fundamental por autocorrelação e para a

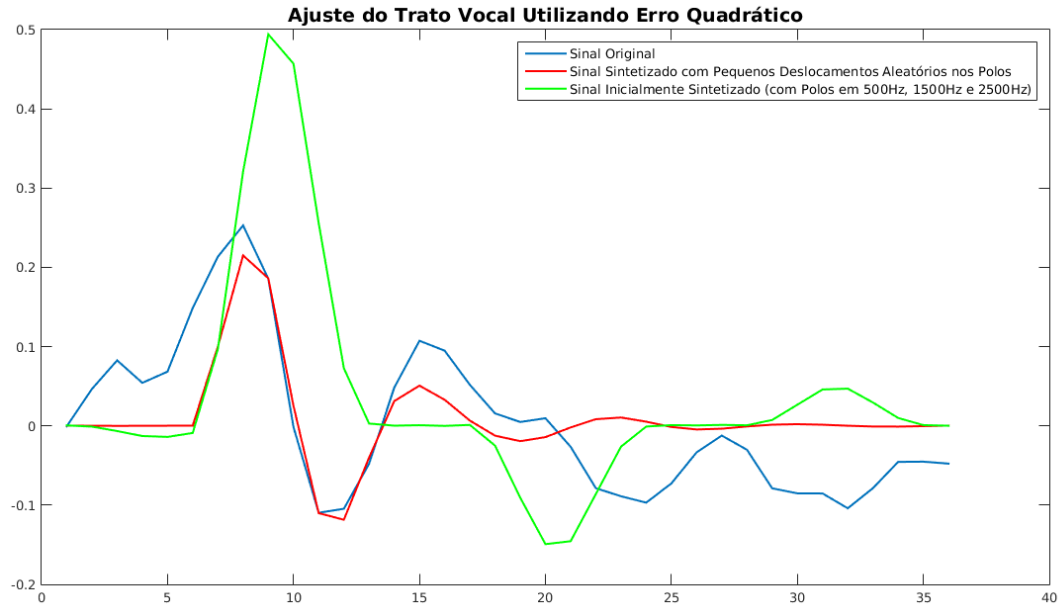

Figura 7. Comparação entre intervalo de um período do sinal original e do sinal sintetizado utilizando o impulso digital como pulso glotal

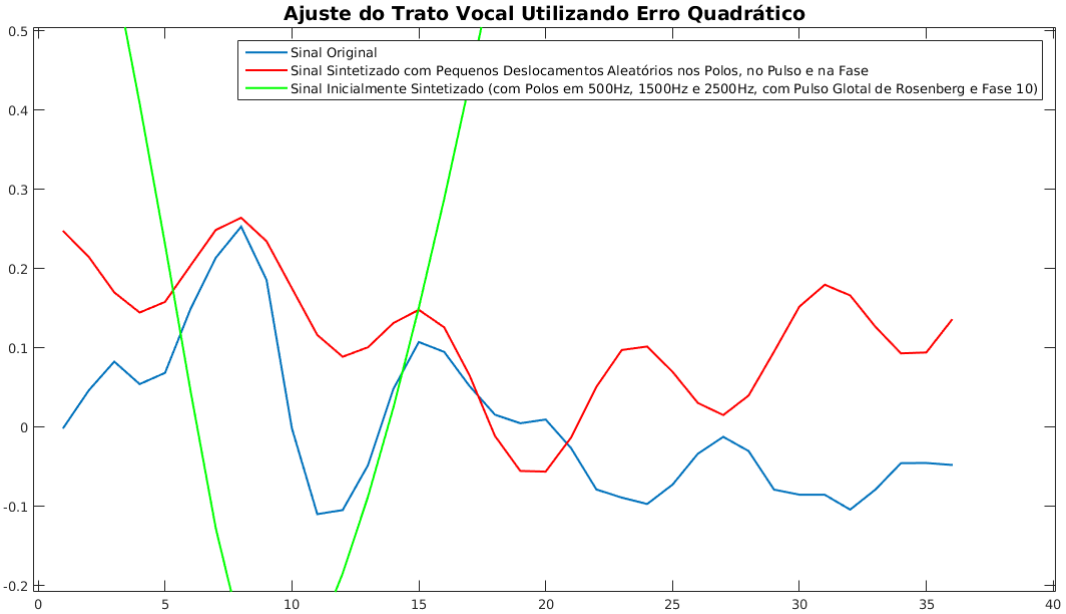

Figura 8. Comparação entre sinal original e sinal sintetizado utilizando pulso glotal de Rosenberg

análise-pela-síntese, denominado $s[n]$, e seu pulso glotal estimado $G[n]$. O IAIF foi implementado utilizando $g = 4$ e $v = 20$ (escolhidos de acordo com a percepção subjetiva do autor), de acordo com o procedimento descrito na subseção 3.2.

Na Figura 11 encontram-se representadas as contribuições espectrais do pulso glotal e trato vocal estimados nas duas iterações realizadas pelo IAIF em azul, e em vermelho o contorno espectral do processo de filtragem inversa. Nela, a estimação final da contribuição do trato vocal é representada por H_{v2}.

Pode ser observado um comportamento coerente com o esperado para o pulso glotal e o trato vocal estimados, de acordo com as representações recorrentes na

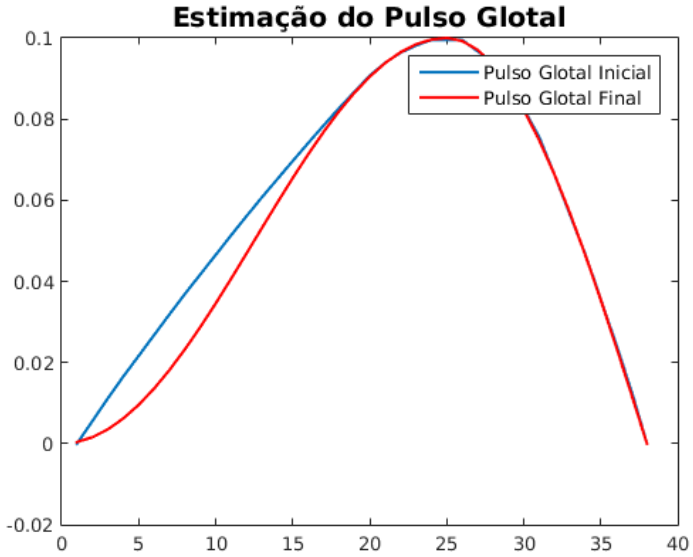

Figura 9. Pulso glotal estimado, a partir do pulso glotal de Rosenberg

literatura. Entretanto, não é possível garantir a acurácia da estimação, sem que haja a comparação com uma representação de referência do sinal glotal. Dessa forma, embora o método IAIF tenha sido estudado e reproduzido detalhadamente neste trabalho de iniciação científica, a validação completa dos resultados para os sinais usados nos experimentos dependeria da disponibilidade de sinais glotais medidos diretamente por métodos invasivos. Por outro lado, os sinais usados no trabalho de referência não estão disponíveis publicamente, o que limitou o aprofundamento desta linha de estudo do trabalho.

Para a avaliação automática da qualidade de voz, utilizando a metodologia proposta na subseção 3.3, foi observado uma maioria de avaliações com rotulações iguais a zero para todas as características (G, R e B). Assim, foi levada em conta também a taxa de ocorrência da avaliação normal para cada característica. Os resultados da classificação linear se encontram representados na Tabela 2. O classificador linear implementado não funcionou adequadamente para a medida R.

Tabela 2. Resultados da Classificação Linear para G e B, utilizando a média e o desvio padrão da f_0 e das formantes

Característica Vocal	Acerto do Classificador	Igual a Zero
G	45.57%	29.11%
B	45.57%	35.44%

Na Figura 12 é apresentado um exemplo da representação facial aplicada em uma amostra de voz classificada com G=0 acompanhada de seu espectrograma. Igualmente, na Figura 13 um exemplo da representação proposta para uma voz classificada com G≠0 e o espectrograma dessa voz são ilustrados.

Partindo da representação espectrográfica, para a voz representada pela face com um sorriso existe uma maior definição das formantes, enquanto que na voz representada pela face "triste" pode se observar uma maior dispersão e uma menor intensidade nas formantes. Até então, avaliações subjetivas feitas pelo autor deste documento e por mais dois voluntários, professores do DEL/UFS, indicam uma

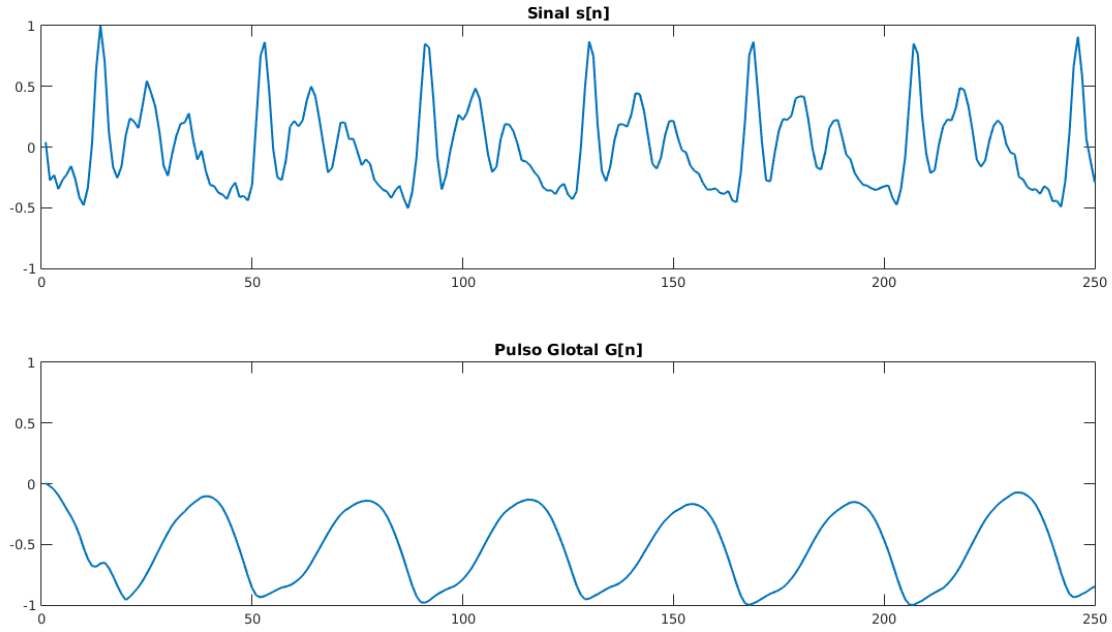

Figura 10. Sinal de voz $s[n]$ e seu pulso glotal estimado $G[n]$

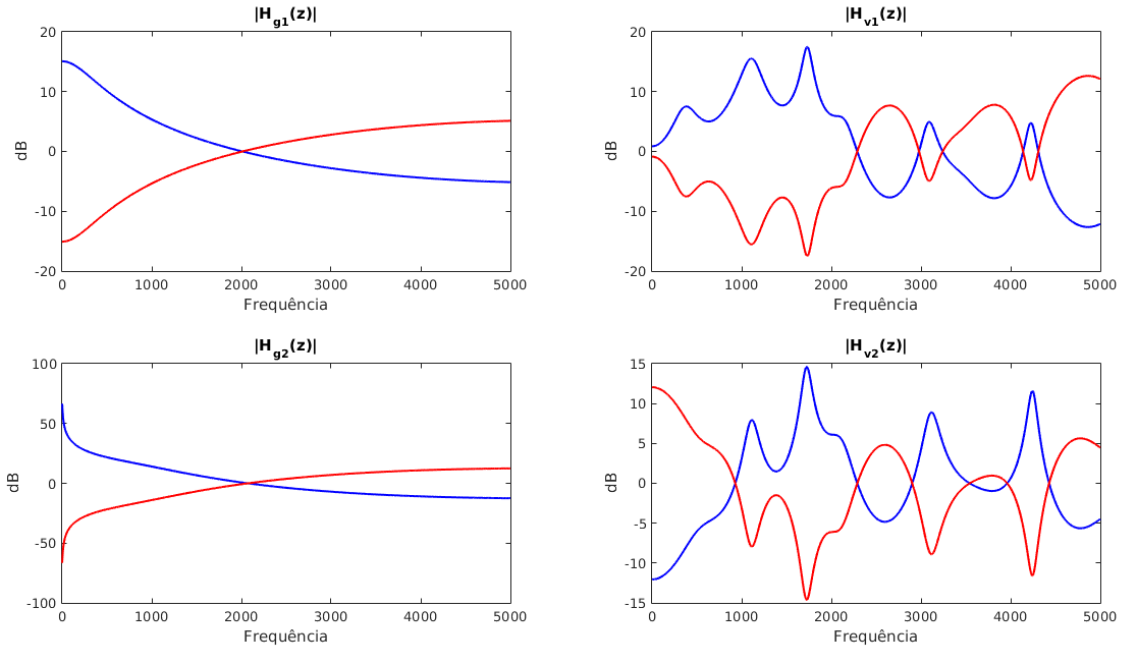

Figura 11. Contornos espectrais estimados para a contribuição glotal e a contribuição do trato vocal, no decorrer das etapas do IAIF

correlação razoável entre o grau da alteração vocal e a distorção percebida na face. Uma investigação detalhada baseada em medidas estatísticas de opiniões (coletadas através de formulários padronizados) será necessária para a consolidação do método de representação.

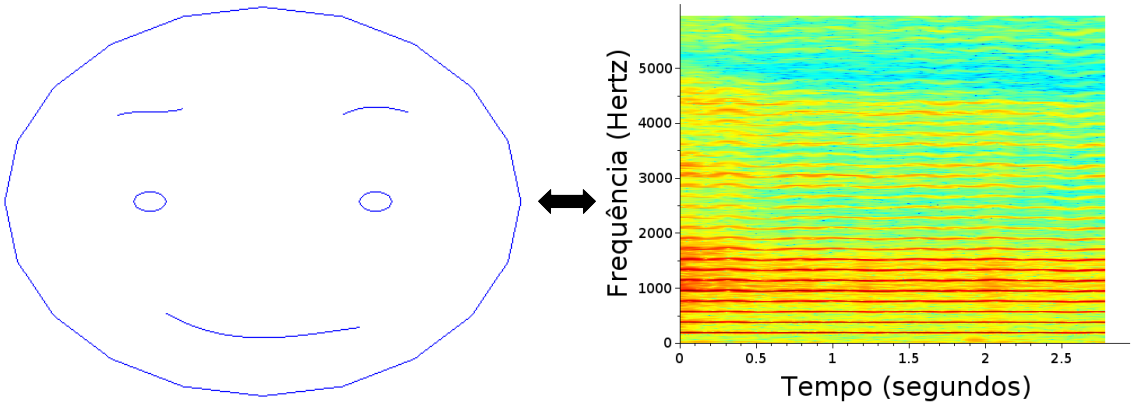

Figura 12. Exemplo da representação da face e espectrograma de uma voz normal (G=0)

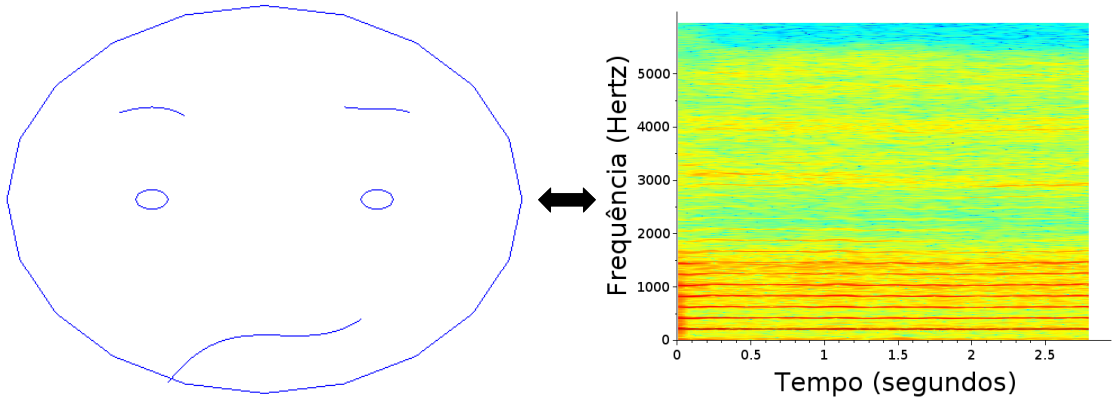

Figura 13. Exemplo da representação da face e espectrograma de uma voz alterada (G≠0)

5 Conclusões

Ao longo do desenvolvimento deste projeto de pesquisa, os diversos objetivos propostos foram trabalhados, de forma que os métodos estudados e implementados proporcionaram um aprendizado significativo do processamento de sinais de fala e reconhecimento de padrões acústicos. As diversas perspectivas abordadas permitem diferentes considerações finais no âmbito do projeto.

Devido à sua simplicidade, o uso da autocorrelação do sinal para a estimação de f_0 não apresenta acurácia em todos os casos, sendo dependente de condições específicas. Devido à necessidade de robustez na estimação em certos casos, são utilizadas com mais frequência versões alteradas da autocorrelação [26] ou outros métodos cujos resultados se apresentaram mais robustos [7][27].No entanto, para a finalidade deste trabalho, em que os sinais vocálicos sustentados favorecem a estimação do f_0, preferiu-se a simplicidade do estimador baseado em correlação na sua versão mais elementar.

Os resultados apresentados pela estimação de pulso glotal e trato vocal utilizando análise-pela-síntese indicam uma metodologia inadequada proposta. Uma investigação mais detalhada é necessária, averiguando as falhas, para que um novo procedimento possa ser proposto. Acredita-se que esse método ainda possa ser adequadamente ajustado para a tarefa de análise da qualidade fonoaudiológica de

sinais, mas esse ajuste não foi obtido na duração deste projeto. Aparentemente, o ajuste dinâmico simultâneo dos parâmetros de pulso glotal e trato vocal sugerem um trabalho investigativo que vai além do escopo (e da duração) de uma iniciação científica, ficando portanto como sugestão de trabalho futuro.

O IAIF, por outro lado, é um método de simples implementação e seus resultados apresentaram-se aparentemente coerentes, de acordo com o referencial teórico. Entretanto, os parâmetros da predição linear são ajustados de maneira subjetiva, sendo necessário um método de ajuste mais objetivo, para garantir melhores resultados. Também se faz necessário realizar a comparação dos sinais glotais estimados com um sinal glotal medido diretamente, para assegurar a acurácia do método, o que não foi possível, para a base de sinais usada.

Os resultados obtidos pela classificação linear indicam que, apesar de uma evidente relação entre os parâmetros e os rótulos das características G e B, a taxa de acertos do classificador é baixa. Isso sugere a possibilidade de que mesmo as medidas de referência feitas por vários especialistas podem discordar, o que levanta questões relevantes sobre a própria utilidade das escalas para o compartilhamento de diagnósticos. Entretanto, também é necessária uma investigação mais completa, analisando as avaliações dos especialistas e experimentando diferentes parâmetros acústicos para a classificação.

Pôde-se observar também que representar informações técnicas relevantes sobre a voz como expressões faciais é uma ferramenta potencialmente útil para a comunicação entre fonoaudiólogo e paciente, que pode ser usada na comunicação intuitiva de exercícios vocais e práticas. Dessa forma, o fonoaudiólogo pode, por exemplo, utilizar o sistema de faces para pedir ao paciente que alcance um rosto 'sorridente' nos exercícios.

6 Perspectivas

Os estudos conduzidos no decorrer do projeto revelam a ampla possibilidade de aplicação dos conceitos estudados no problema investigado. Dessa maneira, além dos problemas e métodos estudados e implementados, pretende-se dar continuidade ao trabalho sob a perspectiva de diferentes aplicações e possibilidades de estudos.

Para uma melhora da estimação da frequência fundamental pretende-se, por meio da investigação da relação entre banco de filtros de banda estreita e predição linear à dois polos, propor um método que estime f_0 de maneira robusta (este trabalho já foi iniciado, junto ao grupo de pesquisa BioChaves/UFS).

Devido aos resultados promissores da implementação do IAIF, pretende-se dar continuidade ao seu estudo. Será realizada, então, a síntese de um sinal de voz, utilizando um pulso glotal e um trato vocal conhecidos, permitindo, assim, a comparação entre o sinal glotal estimado e o utilizado para a síntese. Também serão realizados testes de ajuste do modelo de Liljencrants-Fant [28] à derivada do sinal glotal estimado, onde serão analisados o erro e os parâmetros ajustados, de acordo com metodologia proposta por Strik e Boves [29].

Também é planejado um estudo para analisar estatisticamente a avaliação conjunta de um mesmo conjunto de amostras de voz por parte de grande número de fonoaudiólogos, em escalas perceptivo-auditivas. Para isso, será implementado um programa para coletar avaliações de fonoaudiólogos nas escala GRBASI e CAPE-V. Esse estudo pretende quantificar o nível de concordância entre especialistas nas

escalas estudadas, o que reflete uma questão sensível considerada importante pelos próprios fonoaudiólogos que utilizam e compartilham avaliações graduadas nessas escala.

Para consolidar a representação da voz utilizando desenhos de rostos, pretende-se preparar um estudo baseado em *Mean Opinion Score* sobre a correlação entre avaliação profissional de vozes e percepção das faces correspondentes por leigos.

7 Referências bibliográficas

[1] M. Behlau, *VOZ - O Livro do Especialista.* Rio de Janeiro - RJ: Livraria e Editora Revinter Ltda., 2001, vol. I.

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Prentice-Hall, 1978.

[3] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale," *BioMed Research International*, vol. 2015, no. October, 2015.

[4] A. C. C. Gama, C. F. T. Alves, J. d. S. B. Teixeira, and L. C. Teixeira, "Correlação entre dados perceptivo-auditivos e qualidade de vida em voz de idosas," *Pro Fono*, vol. 21, no. 2, pp. 125–130, 2009.

[5] C. T. Ferrand, *Speech science: an integrated approach to theory and clinical practice.* Allyn & Bacon, 2001.

[6] A. K. Silbergleit, A. F. Johnson, and B. H. Jacobson, "Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality," *Journal of Voice*, vol. 11, no. 2, pp. 222–231, 1997.

[7] O. Babacan, T. Drugman, N. D 'alessandro, N. Henrich, and T. A. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds comparative study of pitch extraction algorithms on a large variety of singing sounds," pp. 1–5, 2013. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00923967

[8] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 140–148, 1973.

[9] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[10] M. M. Sondhi, "New Methods of Pitch Extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[11] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, and T. Dutoit, "A quantitative comparison of glottal closure instant estimation algorithms on a large variety of singing sounds," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1702–1706, 2013.

[12] P. Alku, E. Vilkman, and U. Laine, "Analysis of glottal waveform in different phonation types using the new iaif-method," in *Proc. 12th Int. Congress Phonetic Sciences*, vol. 4, 1991, pp. 362–365.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[14] M. Hunt, J. Bridle, and J. Holmes, "Interactive digital inverse filtering and its relation to linear prediction methods," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78.*, vol. 3. IEEE, 1978, pp. 15–18.

[15] J. H. McClellan, R. W. Schafer, and M. A. Yoder, *Signal processing first*. Pearson education Upper Saddle River, NJ, 2003.

[16] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels." *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[17] D. Paulus and J. Hornegger, *Applied pattern recognition: algorithms and implementation in C++*. Springer Science & Business Media, 2003.

[18] S. H. Friedberg, A. J. Insel, and L. E. Spence, *Linear Algebra: Pearson New International Edition*. Pearson Higher Ed, 2013.

[19] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*. Pearson Upper Saddle River, 2011, vol. 64.

[20] F. L. Wuyts, M. S. De Bodt, and P. H. Van De Heyning, "Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia," *Journal of Voice*, vol. 13, no. 4, pp. 508–517, 1999.

[21] M. Hirano, "Psycho-acoustic evaluation of voice," *Clinical examination of voice*, pp. 81–84, 1981.

[22] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements." *Revue de laryngologie-otologie-rhinologie*, vol. 117, no. 3, pp. 219–224, 1996.

[23] A. Stráník, R. Čmejla, and J. Vokřál, "Acoustic parameters for classification of breathiness in continuous speech according to the GRBAS scale," *Journal of Voice*, vol. 28, no. 5, pp. 653.e9–653.e17, 2014.

[24] R. B. de Alvear, J. Corral, L. Tardón, A. Barbancho, E. Fernández, S. Contreras, A. Martínez-Arquero, and I. Barbancho, "A database and digital signal processing framework for the perceptual analysis of voice quality," in *PAN EUROPEAN VOICE CONFERENCE ABSTRACT BOOK*, 2015, p. 35.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[26] F. Kurth, A. Cornaggia-Urrigshardt, and S. Urrigshardt, "Robust F0 estimation in noisy speech signals using shift autocorrelation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2014, pp. 1468–1472. [Online]. Available: http://ieeexplore.ieee.org/document/6853841/

[27] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, apr 2002. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1458024

[28] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[29] H. Strik and L. Boves, "On the relation between voice source parameters and prosodic features in connected speech." *Speech Communication*, vol. 11, no. 2-3, pp. 167–174, 1992.

8 Outras atividades

No decorrer deste projeto, foi submetido e aprovado um resumo para a apresentação de trabalho, explanando a representação visual proposta para os sinais de voz, no 17º Congresso da FORL[2] (Fundação Otorrinolaringologia). O trabalho será apresentado durante o congresso, nos dias 16 a 18 de agosto de 2018.

Foi escrito e submetido um artigo ao VIII ENCOM - Conferência Nacional em Comunicação, Redes e Segurança da Informação[3]. Nesse artigo, entitulado "Sobre a Associação em Série de Filtros Estreitos e Preditores a Dois Polos - Estudos Preliminares", é estudada a relação entre predição linear e banco de filtros. O resultado da avaliação do artigo está previsto para o dia 13 de agosto de 2018.

Houve, também, a participação nos minicursos da IV SEMAC da UFS entitulados "Redação Científica e Plágio Acadêmico" e "Gerenciamento de Referências Bibliográficas para Trabalhos de Pesquisas e Artigos Científicos".

[2] http://forl.org.br/congresso2018
[3] iecom.org.br/encom2018/